Claims

What is Claimed is:

1. A document processing method for extracting a text block from a document in a document processing system, wherein characters are laid out using blank characters, tabs or any other spaces, comprising the steps of:

generating an object including characters, marks, blank characters and other symbols from the document;

generating a connection candidate between the objects; and

evaluating validity of connection of the connection candidate using a language model.

2. The document processing method according to claim 1, further comprising the steps of:

determining if a connection of the connection candidate is valid; and

combining the objects corresponding to a source and destination of the connection candidate, if it is determined that the connection of the conection candidate is valid.

3. The document processing method according to claim 1, wherein the object generated is associated with a

coordinate indicating a spatial position of the document.

4. The document processing method according to claim 3, wherein the text block is generated by combining the objects, wherein the text block is defined as a rectangular region with a minimum area that includes the objects, wherein the position of the text block is specified by the coordinates of opposing corners of the rectangular region in the document.

5. The document processing method according to claim 1, wherein the connection candidate between the objects is a connection with an object that adjoins the source object on the right side or a connection with an object that is located in the next line and on the left side of the source object.

6. The document processing method according to claim 1, wherein the language model is an N-gram model.

7. The document processing method according to claim 1, wherein the step of generating an object further comprises the steps of:

acquiring a symbol at every spatial coordinate, and associating it with the coordinate in the document;

determining a type of the symbol in a line of the document and generating a token composed of one or consecutive letters, symbols or other characters or a space composed of one or consecutive blank characters;

determining an adjacency relation of the space in the vertical direction and generating a stream composed of spaces that extend over a plurality of lines; and

determining a positional relation between the token and the stream and generating an initial text block that includes the token and blank characters.

8. The document processing method according to claim 7, wherein the step of generating a token or space comprises the steps of:

recording the symbols as a token composed of consecutive characters, if it is determined that the type of symbol is not a blank character and the type of an adjoining symbol in the line is the same;

recording the symbol as one or consecutive spaces if it is determined that the type of symbol is a blank character; and

wherein the step of generating a stream comprises the steps of:

recording the spaces as a stream if it is determined that the spaces are adjoining over different lines in the vertical direction; and

wherein the step of generating an object comprising the steps of:

combining the two tokens and the space between them as an initial text block if a space interposed between two tokens in a line is not a stream.

9. The document processing method according to claim 7, further comprising the steps of:

generating all the initial text blocks and their connection candidates;

extracting initial text blocks of a single element and connection candidates from all the initial text blocks and connection candidates;

determining validity of connection of the initial text blocks of a single element and connection candidates using a language model; and

combining the initial text blocks of a single element if it is determined that the validity of connection is valid.

10. The document processing method according to claim 1, further comprising the steps of:

if there is only a single connection candidate between the objects, the initial text blocks, or the text blocks combined thereof, combining them without determining validity of connection using a language model.

11. A document processing system for extracting a text block from a document, wherein characters are laid out using blank characters, tabs or any other spaces, comprising:

means for generating an object including characters, marks, blank characters and other symbols from the document;

means for generating a connection candidate between the objects; and

means for evaluating validity of connection of the connection candidate using a language model.

12. The document processing system according to claim 11, further comprising:

means for combining the objects corresponding to a source and destination of the connection candidate if it is determined that the connection of the connection candidate is valid.

13. The document processing system according to claim 11, further comprising means for associating a generated object with a coordinate indicating a spatial position of the document.

14. The document processing system according to claim 13, further comprising:

means for generating the text block by combining the

objects, wherein the text block is defined as a rectangular region with a minimum area that includes the objects, wherein the position of the text block is specified by the coordinates of opposing corners of the rectangular region in the document.

15. The document processing system according to claim 11, wherein the connection candidate between the objects is a connection with an object that adjoins the source object on the right side or a connection with an object that is located in the next line and on the left side of the source object.

16. The document processing system according to claim 11, wherein the language model is an N-gram model.

17. The document processing system according to claim 11, wherein the means for generating an object further comprises:

means for acquiring a symbol at every spatial coordinate, and associating the symbol with the coordinate in the document;

means for determining a type of the symbol in a line of the document and generating a token composed of one or consecutive letters, symbols or other characters or a space composed of one or consecutive blank characters;

means for determining an adjacency relation of the space in the vertical direction and generating a

stream composed of spaces that extend over a plurality of lines; and

means for determining a positional relation between the token and the stream and generating an initial text block that includes the token and blank characters.

18. The document processing system according to claim 17, wherein the means for generating a token or space comprises:

means for recording the symbols as a token composed of consecutive characters, if it is determined that the type of the symbol is not a blank character and the type of an adjoining symbol in the line is the same, or for recording the symbol as one or consecutive spaces if it is determined that the type of the symbol is a blank character; and

wherein the means for generating a stream comprises:

means for recording the spaces as a stream if it is determined that the spaces are adjoining over different lines in the vertical direction; and

wherein the means for generating an object comprises:

means for combining the two tokens and the space between them as an initial text block if a space interposed between two tokens in a line is not a

stream.

19. The document processing system according to claim 17, further comprising:

means for generating all the initial text blocks and their connection candidates;

means for extracting initial text blocks of a single element and connection candidates from all the initial text blocks and connection candidates;

means for determining validity of connection of the initial text blocks of a single element and connection candidates using a language model; and

means for combining the initial text blocks of a single element if it is determined that the validity of connection is valid.

20. The document processing system according to claim 11, further comprising:

if there is only a single connection candidate between the objects, the initial text blocks, or the text blocks combined thereof, means for combining them without determining validity of connection using a language model.

21. A computer-readable recording medium having recorded thereon a program for causing a computer to extract a text block from a document, wherein characters are laid out

using blank characters, tabs or any other spaces, the program comprising computer code for causing the computer to perform the steps of:

generating an object composed of characters, marks, blank characters and other symbols from the document;

generating a connection candidate between the objects;

evaluating validity of connection of the connection candidate using a language model; and

if it is determined that the connection of the connection candidate is valid, combining the objects corresponding to a source and destination of the connection candidate.